CREACIÓN AUTOMÁTICA DE SISTEMAS DE BÚSQUEDA DE RESPUESTAS EN DOMINIOS RESTRINGIDOS

Katia Vila, José-Norberto Mazón y Antonio Ferrández

Katia Vila es doctora en ingeniería informática por la *Universidad de Alicante*. Ha sido profesora asistente en el *Departamento de Informática* de la *Universidad de Matanzas* en Cuba, impartiendo docencia en ingeniería informática en asignaturas relacionadas con la inteligencia artificial, la programación descriptiva, y la recuperación de información. Es miembro del *Grupo de investigación en Procesamiento del Lenguaje y Sistemas de Información* (*GPLSI*) del *Departamento de Lenguaje y Sistemas Informáticos* de la *Universidad de Alicante*. Ha publicado diferentes trabajos sobre sistemas de búsqueda de respuestas (BR). Disfruta de una beca *MAEC-AECID* (España) para un proyecto post-doctoral sobre sistemas de BR.

Universidad de Alicante
Departamento de Lenguajes y Sistemas Informáticos
Apartado de Correos 99. 03080 Alicante
kvila@dlsi.ua.es

José-Norberto Mazón es profesor ayudante doctor en el *Departamento de Lenguajes y Sistemas Informáticos* de la *Universidad de Alicante*. Ha publicado varios trabajos en revistas y conferencias especializados de carácter internacional (como *ER*, *ICCS*, *JISBD*, *DSS*, *DKE*, etc.). Ha realizado estancias de investigación en la *Universidad de Münster* (Alemania) y la *Universidad de Toronto* (Canadá). Su investigación se centra principalmente en inteligencia de negocio y desarrollo de software dirigido por modelos. Ha participado en la organización de varias ediciones de los talleres *"Business intelligence and the Web"* (*Beweb*) y *"Web and requirement engineering"* (*WeRE*).

Universidad de Alicante
Departamento de Lenguajes y Sistemas Informáticos
Apartado de Correos, 99. 03080 Alicante
jnmazon@dlsi.ua.es

Antonio Ferrández es profesor en el *Departamento de Lenguajes y Sistemas Informáticos* de la *Universidad de Alicante*. Obtuvo su doctorado en informática en la misma universidad. Sus intereses de investigación son: procesamiento del lenguaje natural, resolución de problemas lingüísticos (anáfora o elipsis), extracción de información, sistemas de recuperación de información y búsqueda de respuestas. Ha participado en proyectos, acuerdos con empresas privadas y organismos públicos relacionados con sus temas de investigación. Ha dirigido tesis doctorales y es autor de artículos de revista y ponencias de congreso.

Universidad de Alicante
Departamento de Lenguajes y Sistemas Informáticos
Apartado de Correos, 99. 03080 Alicante
http://www.dlsi.ua.es/~antonio/antonio.html
antonio@dlsi.ua.es

Resumen

Los sistemas de búsqueda de respuestas (BR) se pueden considerar como potenciales sucesores de los buscadores tradicionales de información en la Web. Para que sean precisos deben adaptarse a dominios concretos mediante el uso de recursos semánticos adecuados. La adaptación no es una tarea trivial, ya que deben integrarse e incorporarse a sistemas de BR existentes varios recursos heterogéneos relacionados con un dominio restringido. Se presenta la herramienta *Maraqa*, cuya novedad radica en el uso de técnicas de ingeniería del software, como el desarrollo dirigido por modelos, para automatizar dicho proceso de adaptación a dominios restringidos. Se ha evaluado *Maraqa* mediante una serie de experimentos (sobre el dominio agrícola) que demuestran su viabilidad, mejorando en un 29,5% la precisión del sistema adaptado.

Palabras clave

Búsqueda de respuestas, Recuperación de información, Dominios restringidos, Desarrollo dirigido por modelos.

Title: Automatic generation of question answering systems in restricted domains

Artículo recibido el 10-11-11
Aceptación definitiva: 11-01-12

Abstract

Question answering (QA) applications can be considered as the potential successors to the traditional information retrieval on the Web. However, QA systems should be adapted to restricted domains for the sake of precision. This adaptation is not a trivial task, since several heterogeneous resources related to a restricted domain must be integrated into existing QA systems. This paper presents the *Maraqa* tool, whose novelty is the use of software engineering techniques such as model driven development to automate the adaptation process. It is worth noting that *Maraqa* has been evaluated through a set of experiments (within the agricultural domain) that demonstrate its applicability: the precision of the adapted QA system showed 29.5% improvement.

Keywords

Question answering, Information retrieval, Restricted domains, Model driven development.

Vila, Katia; **Mazón, José-Norberto**; **Ferrández, Antonio**. "Creación automática de sistemas de búsqueda de respuestas en dominios restringidos". *El profesional de la información*, 2012, enero-febrero, v. 21, n. 1, pp. 16-26.
http://dx.doi.org/10.3145/epi.2012.ene.03

Introducción y motivación

Con la sobreabundancia de información y la heterogeneidad en los formatos de acceso es complicado localizar informaciones concretas. Los usuarios necesitan sistemas que les permitan acceder a información precisa desde los diferentes recursos existentes de una forma transparente y simple.

La primera solución aportada por la comunidad científica para conseguir un acceso sencillo y rápido a la inconmensurable cantidad de información digital accesible, fue la recuperación de información (RI) o *information retrieval* (**Baeza-Yates**; **Ribeiro-Neto**, 1999). La RI consiste en la selección en un depósito de documentos de los que tengan mayor relevancia para una consulta realizada por un usuario.

> "La RI tiene como objetivo obtener respuestas a preguntas precisas formuladas en lenguaje natural"

Los sistemas de RI más conocidos son los que actúan sobre internet y localizan información en la Web, por ejemplo los motores de búsqueda como *Google* o *Yahoo*. Según **Russell** y **Norvig** (2003), este modelo de RI corresponde casi totalmente a palabras, ya que acepta una cantidad mínima de sintaxis (palabras que deben aparecer una junto a la otra), y un papel diminuto de clases semánticas (en forma de listas de sinónimos). Por ello los resultados son en muchas ocasiones documentos que contienen los términos de la consulta, pero no son la información deseada. Por otro lado, la salida es una lista de documentos ordenada en función de medidas de similitud con la pregunta y medidas indirectas de la visibilidad y prestigio de la información (por ejemplo *page rank* de la web), lo cual puede convertir a esos motores en sistemas manipulables. Luego resta una ardua tarea ya que el usuario debe revisar y leer cada documento de la lista obtenida, lo primero para ver si en realidad está relacionado con los requerimientos solicitados y lo segundo para localizar en su interior la información puntual que se desea.
http://www.google.com
http://www.yahoo.com

Los inconvenientes mencionados impulsaron la investigación en sistemas de búsqueda de respuesta (BR) o *question answering*, que tienen como objetivo la obtención de respuestas concretas a preguntas precisas indicadas por el usuario directamente en lenguaje natural.

En las figuras 1 y 2 se muestran las salidas para la pregunta *What year did Wilt Chamberlain score 100 points?* de un sistema de RI como *Google* y de un sistema de BR como *NSIR* respectivamente. En el caso de la RI se aprecia, en contraste con la BR, que el usuario ha de buscar dentro de los documentos devueltos.
http://clair.si.umich.edu/clair/NSIR/html/nsir.cgi

Los sistemas de BR deben adquirir un nivel de comprensión del texto muy superior al alcanzado por la RI, por lo que es habitual realizar un análisis léxico, sintáctico y semántico tanto de la pregunta como de los documentos. Este tipo de análisis conlleva un coste computacional muy elevado, el cual se supera mediante el filtrado de los documentos a través de la RI, es decir, entre los millones de documentos de partida, la BR realiza el análisis y búsqueda de la respuesta únicamente sobre unos cientos de pasajes (conjunto de oraciones de un documento).

Por último, se analizan las diferencias entre BR y RI desde el punto de vista de la propia pregunta que realiza el usuario. En primer lugar, habitualmente la pregunta en RI no es más que una serie de palabras ("invención teléfono"), mientras que en BR ésta ha de ser una pregunta formulada correctamente en lenguaje natural ("¿Cuándo fue inventado el teléfono?"). En segundo lugar, la RI se centra en la propia pregunta, mientras que la BR se centra en la respuesta a dicha pregunta. Por ejemplo, si planteamos la pregunta "¿Cuándo fue inventado el teléfono?" a un sistema de RI, buscará el documento que más veces contenga "teléfono", mientras que la BR buscará el que contenga la fecha concreta de la invención del teléfono.

La investigación en BR se ha visto incentivada a partir de 1999 por conferencias como *TREC* (*Text retrieval conference*) desde su octava edición, *CLEF* (*Cross-language evaluation forum*) y *Ntcir* (*NII Test collection for IR systems*). En estas conferencias compiten sistemas de BR desarrollados tanto por instituciones académicas como por empresas. Su

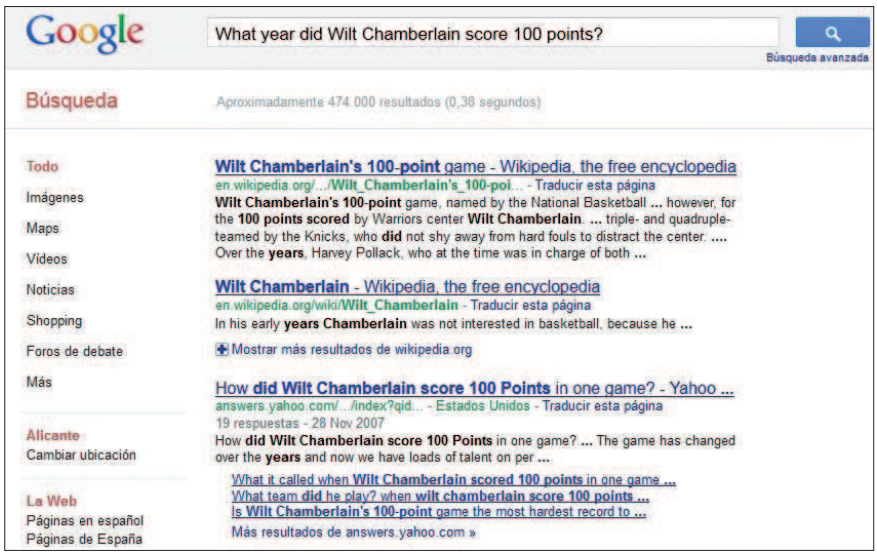

Figura 1. Salida de un sistema de RI para una pregunta en lenguaje natural

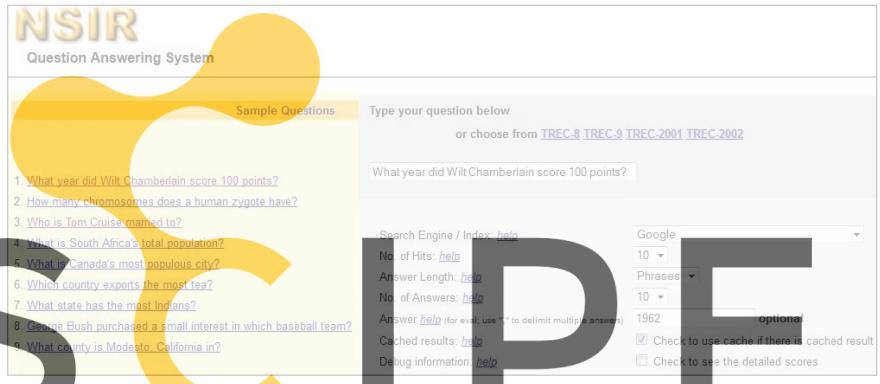

Figura 2. Salida de un sistema de BR para una pregunta en lenguaje natural

funcionamiento general (dependiendo del tipo de subtarea sobre la que se compite) consiste en que los participantes disponen de unas colecciones de documentos y una serie de preguntas, cuya respuesta es analizada por un comité que ordena cada sistema por una serie de medidas. Estas colecciones de documentos suelen ser de un tamaño considerable y de contenidos diversos conocidos como de "dominio abierto" (DA).
http://trec.nist.gov
http://www.clef-initiative.eu
http://research.nii.ac.jp/ntcir/index-en.html

Algunos ejemplos de sistemas de BR con una dilatada trayectoria son *Start* (*SynTactic analysis using reversible transformations*) y *Wolfram Alpha*.

Start, operativo desde diciembre de 1993, fue realizado por **Boris Katz** y los componentes del *InfoLab Group* del *MIT Computer Science and Artificial Intelligence Laboratory* (**Katz**, 1997). Está basado en el etiquetado previo de segmentos de texto según diferentes granularidades (**Katz** *et al.*, 2006), lo cual facilita el acceso a los tipos de respuesta predefinidos, pero presenta la desventaja de limitar los tipos de pregunta a realizar al sistema.
http://start.csail.mit.edu

Siguiendo la misma estrategia de *Start*, *Wolfram Alpha* (anunciado en marzo de 2009) –de la compañía *Wolfram Research*, presidida por **Stephen Wolfram**–, extrae respuestas de una base de datos estructurada. Este sistema indexa una gran cantidad de respuestas, sobre las cuales realiza un proceso de inferencia cuando se plantea una pregunta. Es decir, ha construido una gran base de datos de conocimiento que utiliza para este proceso de inferencia. De un modo similar funciona *Cyc*, el cual aúna una ontología y una base de datos de conocimiento general para llevar a cabo razonamientos de tipo *humano*. Este tipo de planteamientos adoptados tradicionalmente tienen limitaciones de extensibilidad debido a la estructuración predefinida de los tipos de respuestas.
http://www.wolframalpha.com
http://www.cyc.com

A pesar de los amplios esfuerzos investigadores realizados, las carencias de los sistemas de RI actuales y el acuerdo general de que los sistemas de BR son sus potenciales sucesores, no se ha generalizado su uso. El motivo principal es que la BR no es lo suficientemente precisa sobre un dominio abierto para el coste computacional que supone, siendo necesarios muchos recursos semánticos y una capacidad de razonamiento no alcanzada de momento.

Estos objetivos sí se podrían alcanzar sobre "dominios restringidos" (DR) tales como el dominio médico o legal, en los que se disponga de recursos semánticos suficientes.

> "A pesar de las investigaciones, las carencias de los sistemas de RI y la creencia de que los sistemas de BR son sus sucesores, no se ha generalizado su uso"

Nuestro trabajo se centra en la adaptación automática de sistemas de BR en DA a dominios restringidos. Se presenta la herramienta *Maraqa* (*Model-driven adaptation for restricted-domain question answering*), que automatiza el proceso de adaptación y por tanto, puede usarse como un recurso que otorgará a todo sistema de BR en DA la capacidad para enfrentar el cambio, para introducir y extraer información de diferentes dominios restringidos independientemente del idioma y formato de sus recursos de conocimiento. Para ello, nos basaremos en el desarrollo dirigido por modelos o *model-driven development* (**Bézivin**, 2005).

Arquitectura de un sistema de BR

Después de estudiar varias de las aproximaciones de BR presentadas en conferencias como *TREC* y *CLEF*, se ha observado que la mayoría de ellas tiene una arquitectura común que se muestra en la figura 3.

El módulo de indexación incluye los procesos aplicados a la colección de documentos de manera offline con el objetivo de acelerar el proceso de BR:

– indexación de documentos para el sistema de RI, que usa información estadística (se asignan pesos a las palabras);
– indexación de documentos para el proceso de BR, que usa información léxica, sintáctica y semántica.

Por otro lado el módulo de búsqueda es el encargado de realizar 3 procesos para responder a la pregunta del usuario:

1. Análisis de la pregunta, en el cual
   – se clasifica la pregunta [ej.: en función del tipo de respuesta esperada (TRE), que nos indica el tipo semántico de la información a buscar, como "tipo persona" para la pregunta "¿Quién es el presidente de España?"];
   – se extraen las palabras clave de la pregunta para usarlas como consulta en el siguiente proceso de RI (ej.: para "¿Cuál es el correo del *Ministerio de Economía y Competitividad*?", es habitual descartar el término "correo" porque éste no suele aparecer en la respuesta y su inclusión en la pregunta de RI podría devolver pasajes que contengan dicho término en lugar de la información buscada. Esto es lo que ocurre en la figura 1, en la que se muestran pasajes con las apariciones del término *year* para la pregunta *What year did Wilt Chamberlain score 100 points?*).
2. Recuperación de documentos o pasajes relevantes, donde se utilizan las palabras clave previamente detectadas y el indexado de los documentos para la RI realizado con anterioridad, como entradas del sistema de RI. La salida serán los documentos o pasajes relevantes para la pregunta del usuario, reduciendo así el espacio de búsqueda para la BR.
3. Análisis del conjunto de documentos o pasajes relevantes usando el indexado para la BR previamente realizado, con la finalidad de encontrar la respuesta esperada por el usuario.

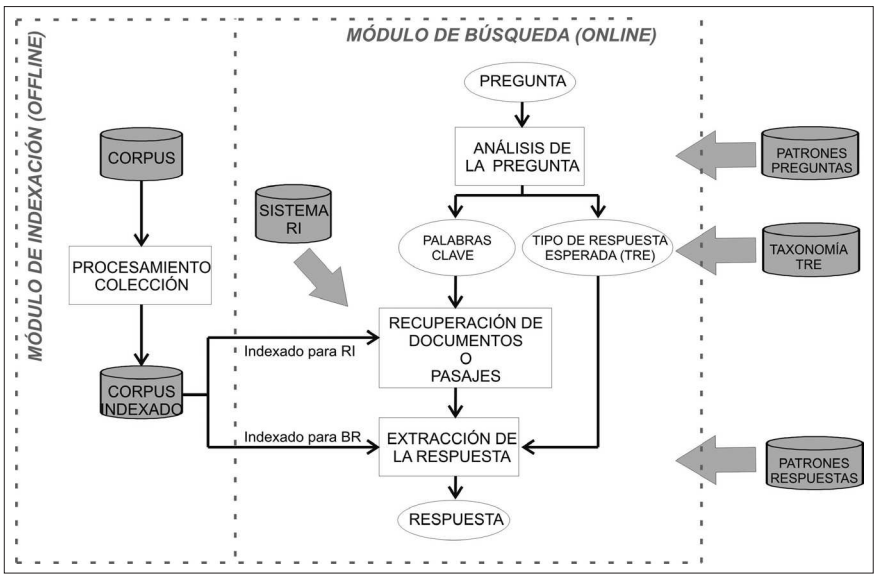

Figura 3. Arquitectura típica de un sistema de BR

Problemas en la adaptación de un sistema de BR de DA a DR

Si analizamos la arquitectura de un sistema de BR es obvio que las fases de análisis de la pregunta y extracción de la respuesta son dependientes del conocimiento, usualmente incluido en patrones, que se tenga del dominio de aplicación. En el contexto de este trabajo, se entiende por "patrones" (**Vila**, 2010) todas las posibles estrategias para la detección de las relaciones entre los elementos de la pregunta y la respuesta (ej.: formas lógicas, expresiones regulares, relaciones sintácticas, relaciones de dependencia, etc.). Además se puede destacar la repercusión que tiene la fase de análisis de la pregunta en el resto de fases de la arquitectura típica de un sistema de BR y por tanto, en la eficiencia del sistema.

En esta fase es crucial la determinación del tipo semántico de la respuesta que espera el usuario o TRE, por medio de una taxonomía de TRE [también conocida como jerarquía de preguntas (**Li**; **Roth**, 2006) u ontología de preguntas (**Metzler**; **Croft**, 2005)] predefinida. Una correcta especificación de la taxonomía de TRE implica una efectiva detección del TRE de la pregunta del usuario, reduciendo así el espacio de búsqueda de las respuestas candidatas y brindando una respuesta más precisa (**Li**; **Roth**, 2006; **Hovy**; **Hermjakob**; **Ravichandran**, 2002). Un ejemplo ilustrativo sería para la pregunta del usuario "¿Quién es el presidente de España?", el TRE sería "persona" y las respuestas candidatas serían nombres propios como "Mariano Rajoy".

> "Para el funcionamiento de un sistema de BR de DA en dominios restringidos es necesaria la adaptación de sus patrones de preguntas y respuestas y de la taxonomía de TRE que se utilice"

Teniendo en cuenta estas consideraciones, se puede afirmar que para lograr el correcto funcionamiento de un sistema de BR de DA en dominios restringidos es necesaria la adaptación de:

– los patrones de preguntas y respuestas del sistema de BR;
– la taxonomía de TRE que se utilice.

Sin embargo, en las aproximaciones actuales para alcanzar dicha adaptación persisten dos problemas fundamentales:

– adaptación manual de los patrones de BR (**Ferrés**; **Rodríguez**, 2006; **Peñas** *et al.*, 2009) y las taxonomías de TRE (**Sekine**; **Sudo**; **Nobata**, 2002; **Hovy**; **Hermjakob**; **Ravichandran**, 2002; **Metzler**; **Croft**, 2005; **Li**; **Roth**, 2006) a dominio restringidos, lo cual requiere de un esfuerzo enorme en función del tiempo y del coste propios de la complejidad inherente de los conceptos provenientes de esos dominios (**Mollá**; **Vicedo**, 2007);

– definición de los patrones de BR y las taxonomías de TRE para dominios restringidos por medio del análisis de corpus de preguntas-respuestas potenciales (**Terol**; **Martínez-Barco**; **Palomar**, 2006; **Kosseim**; **Yousefi**, 2008), lo que no es una situación realista ya que es muy difícil y compleja la adquisición de este tipo de corpus en dominios restringidos.

Después de analizar la situación actual de adaptación de sistemas de BR en DA a dominios restringidos, parece indiscutible la necesidad de diseñar estrategias que faciliten este proceso. Principalmente, las propuestas deben:

– elevar el grado de automatización de la adaptación, evitando que sea un proceso tedioso y complejo;
– explotar los recursos de conocimiento disponibles para un dominio concreto independientemente de su esquema de representación;
– utilizar el corpus textual como punto de partida y fuente de información principal, en detrimento del (poco realista en dominios restringidos) uso de corpus de preguntas;
– aumentar la precisión de la BR.

Para alcanzar estos objetivos, presentamos *Maraqa*, una herramienta que permite una adaptación sistemática de los sistemas de BR en DA a nuevos dominios restringidos basada en el paradigma de ingeniería del software conocido como desarrollo dirigido por modelos (*model driven development*, *MDD*), que se basan en la representación de los conceptos y actividades que rigen un área concreta de conocimiento en lugar de usar terminología informática o codificación de algoritmos en lenguajes de programación complejos (**Mellor** *et al.*, 2003). La creación de estos modelos debe basarse en una sintaxis gráfica, mediante reglas y normas establecidas en un metamodelo (**Kleppe** *et al.*, 2003). Además, mediante el uso de *MDD* se posibilita la generación automática de la aplicación software a partir de modelos (**Mellor**; **Clark**; **Futagami**, 2003). Por lo tanto, *MDD* enfatiza dos aspectos clave: los modelos y las transformaciones entre ellos para llegar a obtener el código fuente correspondiente del sistema software.

Maraqa

Ha sido desarrollado siguiendo el paradigma *MDD* mediante la plataforma *Eclipse*. Permite adaptar fácilmente sistemas de BR en DA a dominios restringidos, haciendo a estos sistemas útiles para su uso en dominios técnicos y no triviales que necesitan de una alta precisión como medicina, farmacia, derecho, etc. Por ejemplo, en el dominio agrícola una pregunta tipo podría ser "¿Qué enzima incrementa la digestibilidad del fósforo orgánico en los animales?". La pregunta es sobre un dominio restringido por lo que su terminología es más específica y por lo tanto es incorrectamente respondida por sistemas de BR en DA. En el trabajo de **Vila** *et al.* (2011) se realizó una serie de experimentos para comparar la precisión de un sistema de BR de DA desarrollado en la *Universidad de Alicante*, llamado *AliQAn* (**Roger** *et al.*, 2008) y su adaptación al dominio agrícola mediante *Maraqa*. Los resultados indicaron que la precisión del sistema de BR *AliQAn* fue de 28,8% comparado con la media del sistema en un dominio abierto que es alrededor del 43% de precisión. Una vez *AliQAn* fue adaptado usando *Maraqa*, la precisión alcanzó un 58,3% (**Vila**; **Mazón**; **Ferrández**, 2011). En el apartado de evaluación de *Maraqa* se describen con mayor profundidad los experimentos realizados.
http://www.eclipse.org

> *Maraqa* permite adaptar sistemas de BR en dominios abiertos a dominios restringidos

Para adaptar sistemas de BR a DR, *Maraqa* requiere los siguientes recursos:

a) Conocimiento no estructurado: en este caso se refiere a toda la documentación no estructurada disponible en el dominio restringido al cual se desea adaptar el sistema de BR. Esta documentación es conocida como colección de documentos o corpus.
b) Conocimiento semi-estructurado o estructurado: cualquier sistema de organización del conocimiento (SOC) –llamados en inglés *knowledge organization systems*, *KOS* (**Hodge**, 2000)–, disponible en el dominio. Estos recursos de conocimiento incluyen una variedad de esquemas para organizar, manejar y recuperar información. El término SOC pretende abarcar cualquier tipo de esquema que sirva para gestionar el conocimiento, como diccionarios, taxonomías, tesauros, ontologías, etc. Según las áreas de conocimiento a las que se refiera un SOC, existen dos tipos: SOC genérico como *WordNet*, *EuroWordNet*, *SUMO*; o SOC de dominio, tales como el tesauro *Agrovoc* para el dominio agrícola, el tesauro multilingüe *EuroVoc* que abarca la terminología de los ámbitos de actividad de la Unión Europea, el metatesauro *UMLS* para el dominio médico, etc.
http://wordnet.princeton.edu
http://www.illc.uva.nl/EuroWordNet
http://www.ontologyportal.org
http://www.fao.org/agrovoc
http://eurovoc.europa.eu
http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus
c) Código fuente de los patrones de preguntas y respuestas del sistema de BR en DA que se desea adaptar.

Cabe resaltar que *Maraqa* es independiente del idioma de los recursos que emplee, pero para su correcto funcionamiento todos estos recursos deben estar en el mismo idioma.

A partir de estos recursos, *Maraqa* obtiene:

– taxonomía de TRE refinada para el dominio en cuestión;
– código fuente de los patrones de preguntas adaptados al dominio;
– código fuente de los patrones de respuesta adaptados al dominio.

Para explicar la herramienta utilizaremos como caso de estudio la pregunta "¿Qué glicósidos tienen un efecto defaunante en el rumen?". Nuestro objetivo sería adaptar *AliQAn* a este dominio agrícola. Para ello se necesita, en primer lugar, un corpus o conocimiento no estructurado, sobre el

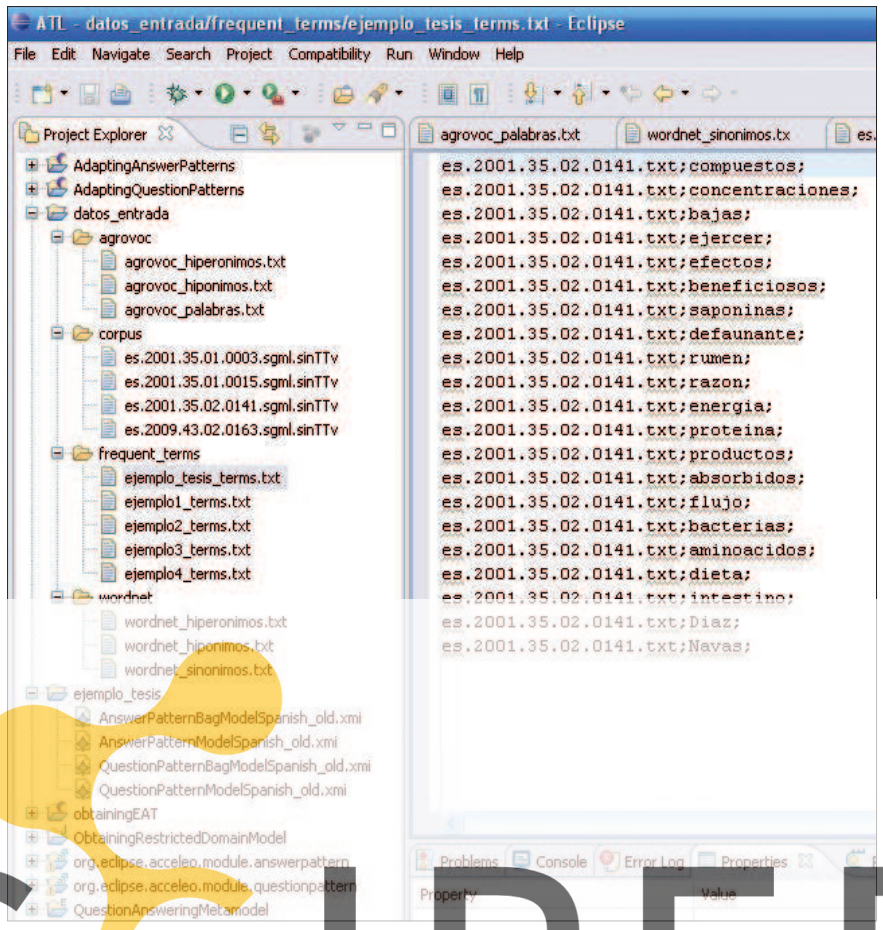

Figura 4. Obtención de términos más relevantes

Figura 5. Ejemplo de relaciones sintácticas entre términos relevantes.

dominio. En este caso nuestro corpus es un conjunto de artículos de investigación de la *Revista cubana de ciencia agrícola* (*RCCA*). En este corpus se encuentra el fragmento de texto que contiene una respuesta candidata: "[…] Sin embargo, en ocasiones, estos compuestos, en concentraciones bajas, pueden ejercer efectos beneficiosos. Por ejemplo, las saponinas tienen un efecto defaunante en el rumen, lo que puede contribuir al aumento de la razón energía: proteína de los productos absorbidos debido al incremento del flujo de bacterias y de aminoácidos de la dieta hacia el intestino […]".
http://www.ica.inf.cu/revista-cubana-de-ciencia-agricola/index.php
http://redalyc.uaemex.mx/src/

El primer paso realizado por *Maraqa* consiste en determinar los términos más relevantes del dominio a partir de la indexación del corpus (para RI y BR según la figura 3), tomando los términos que sean sustantivos, verbos y adjetivos y que tengan una determinada frecuencia de aparición en el corpus. Para nuestro caso, los términos más relevantes se pueden ver en la captura de pantalla de *Maraqa* de la figura 4 precedidos por el documento en el que aparecen (ej.: "compuestos" que aparece en el documento "es.2001.35.02.0141.txt").

Se obtienen las relaciones entre estos términos más relevantes usando el análisis sintáctico del corpus realizado con anterioridad en el indexado para BR (ver fase offline de figura 3); por medio de herramientas de procesamiento del lenguaje natural: primero se utilizó el etiquetador léxico o *PoS tagger MACO* (**Acebo** *et al*., 1994) y luego el analizador sintáctico *Supar* (**Ferrández**; **Palomar**; **Moreno**, 1999). Por ejemplo, si un sustantivo puede ser sujeto de un verbo o qué adjetivos modifican a un sustantivo. En la figura 5 se observa un ejemplo de este análisis con *Maraqa* (la relación adjetivo-sustantivo entre los términos "importante-papel", o la relación sujeto-verbo entre "biota-desempeña").

Además, se debe tener en cuenta el conocimiento estructurado tanto del dominio agrícola como del dominio abierto. En este caso para el primero *Maraqa* usa el tesauro *Agrovoc* (en la figura 6 aparece cada término del te-

sauro precedido por su identificador), mientras que se usa *WordNet* como recurso de dominio abierto (figura 7).

Luego *Maraqa* debe obtener los patrones existentes de *AliQAn* tanto de pregunta (figura 8) como de respuesta (figura 9), con el fin de adaptarlos al dominio agrícola. En la figura 8 se aprecia cómo se extrae la estructura del patrón de pregunta "patronEO5" con:

– las expresiones "ptdtEO5" (con valor "qué") y "snsEO5", relacionado con el concepto "objeto_inanimado" y todos sus hipónimos, como "edificio" o "instrumento_musical";
– la asociación "ptdtEO5-snsEO5" que sirve para relacionar ambas expresiones;
– el TRE "entidad_objeto" que sirve para identificar preguntas del tipo: "¿Qué instrumento musical tocaba Beethoven?"

En la figura 9 se aprecia cómo se extrae la estructura del patrón de respuesta "patronRespEO" con:

– TRE "entidad_objeto";
– los conceptos asociados de "objeto_inanimado" y sus hipónimos; el cual permite encontrar la respuesta a la pregunta anterior en el siguiente texto "[…] Beethoven, último gran representante del clasicismo vienés, tocaba el piano desde joven […]".

Una vez se tienen todos estos recursos, *Maraqa* obtiene un modelo del dominio restringido donde se determina la equivalencia entre los términos del corpus y los conceptos del dominio restringido. Después obtiene todos los conceptos relacionados usando SOC de dominios y genéricos. La figura 10 muestra el modelo generado por *Maraqa* donde se observa que el término "saponinas" tiene un concepto equivalente en el SOC de dominio (*Agrovoc*) con código "6795". Luego se obtienen los conceptos relacionados con el mismo en *Agrovoc* (como los hipónimos "glicoalcaloides" y "ginsenósidos" y los hiperónimos "glicósidos", "carbohidratos" y "compuestos_organicos") y en *WordNet* (como los hiperónimos "compuesto_químico", "sustancia", "objeto_inanimado" y "entidad").

Una vez obtenido el modelo de dominio restringido, *Maraqa* genera una taxonomía de TRE propia del dominio

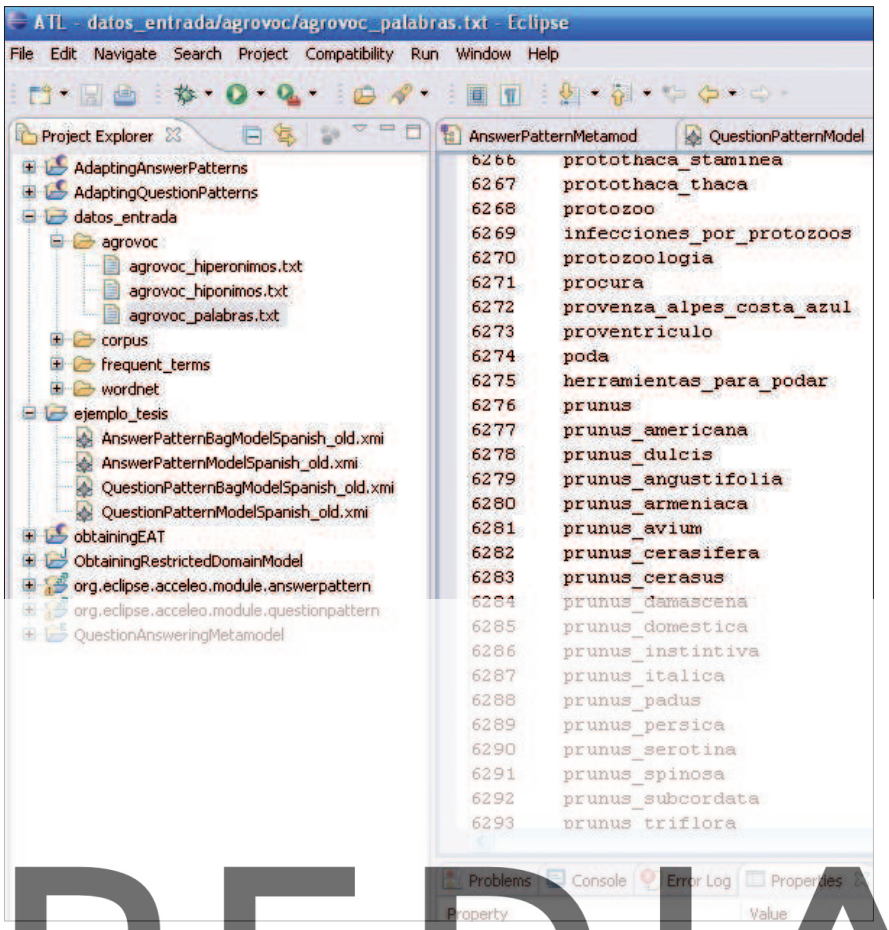

Figura 6. Ejemplo del uso de *Agrovoc* por *Maraqa*

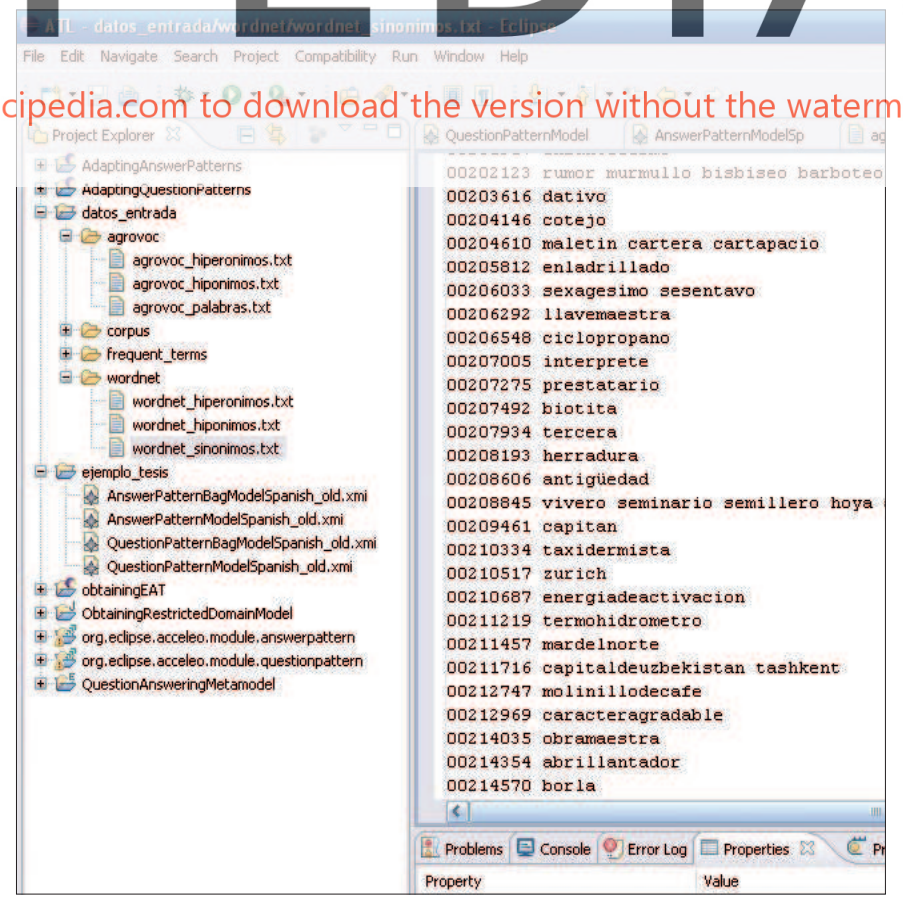

Figura 7. Ejemplo del uso de *WordNet* por *Maraqa*

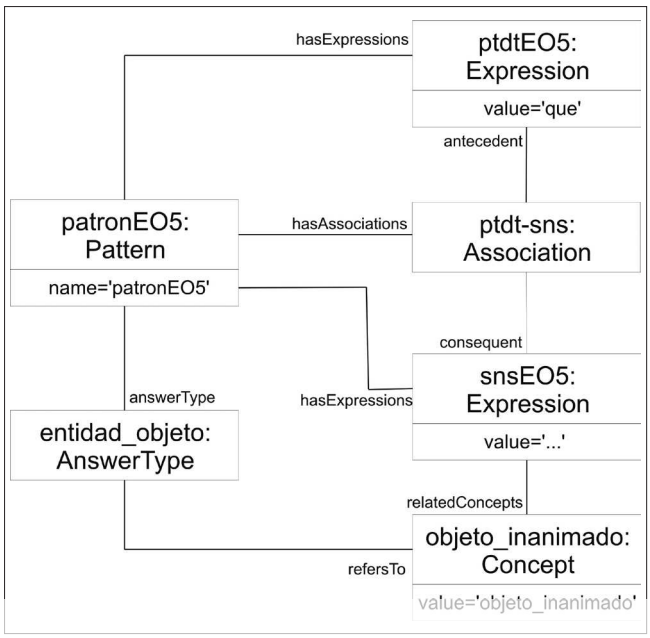

Figura 8. Ejemplo de patrones de pregunta de DA extraídos por *Maraqa*

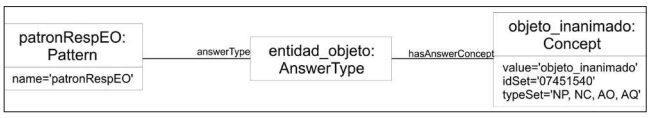

Figura 9. Ejemplo de patrones de respuesta de DA extraídos por *Maraqa*

restringido que estamos tratando. La figura 11 muestra la correspondiente para nuestro caso de estudio, donde se observa que el concepto "glicósidos" tiene como hiperónimo a "compuesto orgánico". Siguiendo esta taxonomía, la pregunta ejemplo tendría como TRE "glicósidos" y el espacio de búsqueda estaría sólo restringido a tipos de "glicósidos" (como "saponinas") los cuales podían ser aceptados como respuestas correctas.

Finalmente, con la taxonomía de TRE para el dominio específico, *Maraqa* adapta los patrones de pregunta y respuesta existentes al dominio (en este caso agrícola) y genera el código correspondiente a los mismos para ser adicionado al componente de BR de [illegible] de pregunta adaptado se muestra en la figura 12 y su correspondiente código en *C++* en la figura 13. Dicho patrón permite dar respuesta a la pregunta que se ha empleado de ejemplo "¿Qué glicósidos tienen un efecto defaunante en el rumen?", la cual sería "saponinas"

En la siguiente sección explicaremos cómo fueron realizados los experimentos y expondremos los resultados obtenidos.

Evaluación de *Maraqa* en el dominio agrícola

Para realizar la evaluación empleamos los primeros 43 volúmenes de la *Revista cubana de ciencia agrícola* (*RCCA*) como dominio restringido agrícola. Es una publicación del *Instituto de Ciencia Animal*, perteneciente al *Ministerio de Educación Superior* de la República de Cuba, editada en español e inglés desde 1967. Cada uno de los volúmenes tiene un promedio de tres o cuatro números, lo que hace un total de 140 números y 2.000 artículos (28,65 MB como ficheros pdf). En concreto, el corpus textual *RCCA* está formado por un conjunto de ficheros de texto plano creados a partir de los archivos originales (con formato de texto enriquecido) de la *RCCA*, usando la utilidad de *Linux pdftotext*. De esta manera se evitaron los problemas con la codificación de los caracteres o con la maquetación (formato de doble columna, encabezados y pies de página, conversión de tablas y fórmulas, etc.). Para llevar a cabo los experimentos se contó con un conjunto de 330 preguntas *RCCA* elaboradas por expertos del dominio. La inmensa mayoría de las preguntas son de tipo factual y están basadas en hechos, preguntando por nombres de plantas, microorganismos, sustancias, fármacos, compuestos químicos, alimentos, fauna, personas, localización, día en el que sucedió algún hecho, etc. Algunos ejemplos de preguntas son: "¿Qué es la necrosis cerebrocortical?" o "¿Qué produce la cytophaga?".

Después de usar la herramienta *Maraqa* y los recursos de conocimiento disponibles (corpus *RCCA*, *Agrovoc* y *WordNet*), se obtuvo un total de 9.022 términos relevantes, 921 conceptos en la taxonomía de TRE para el dominio agrícola de la *RCCA*, y 2.600 y 325 patrones de preguntas y respuestas, respectivamente. De esta manera obtuvimos el sistema de BR para el dominio agrícola de la *RCCA* (**Vila**; **Mazón**; **Ferrández**, 2011).

Para verificar la eficiencia del sistema de BR para el dominio agrícola, obtenido a partir de *AliQAn* y usando *Maraqa*, se llevaron a cabo dos experimentos (**Vila**, 2010; **Vila**; **Mazón**; **Ferrández**, 2011). Ambos tenían como objetivo medir la precisión del sistema de BR de DA [illegible] dominio agrícola. El primer experimento usando el sistema de BR de DA *AliQAn* (SBR-DA *AliQAn*) y el segundo empleando el sistema de BR de DR agrícola (SBR-DR *Agrícola*). Los resultados de ambos experimentos (figura 14) demostraron que la precisión del sistema de BR de DR *agrícola* era mayor que la precisión

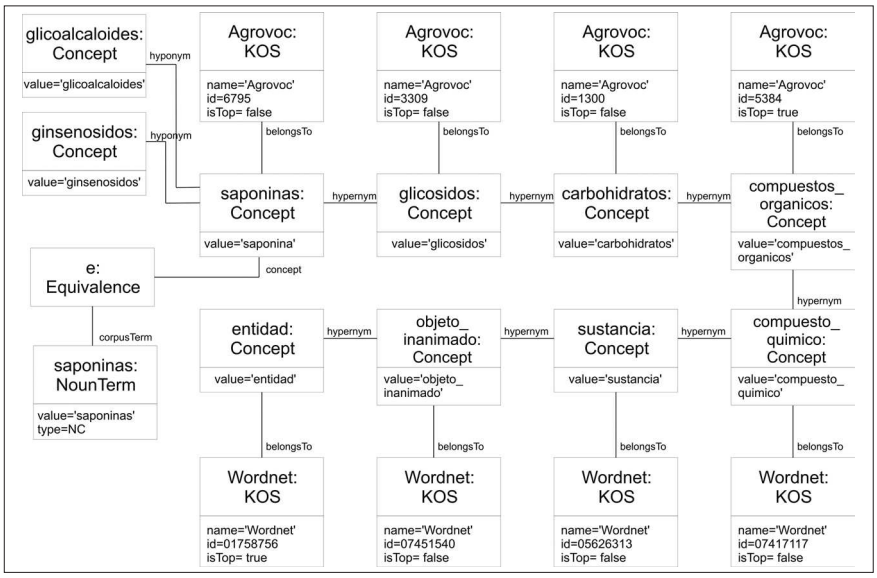

Figura 10. Ejemplo del modelo de dominio restringido y representación del concepto "saponinas" por *Maraqa*

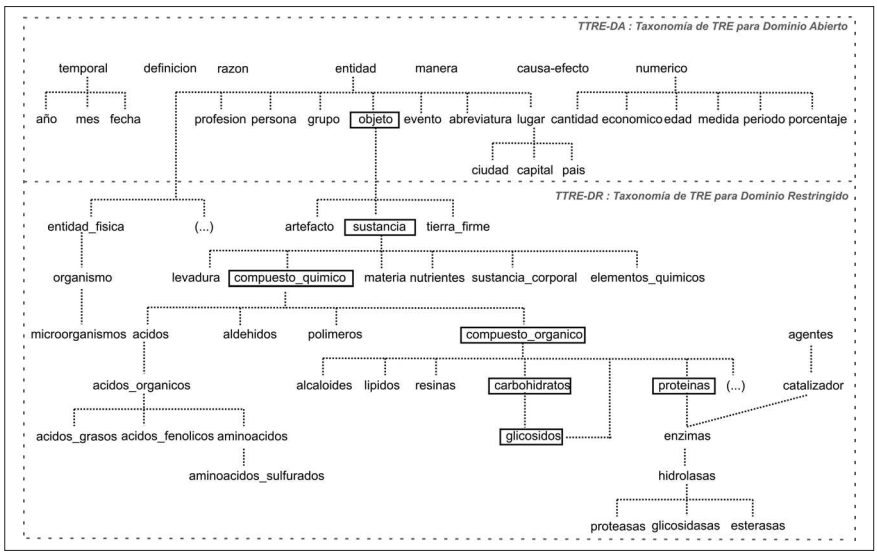

Figura 11. Ejemplo de taxonomía de TRE obtenida por *Maraqa*

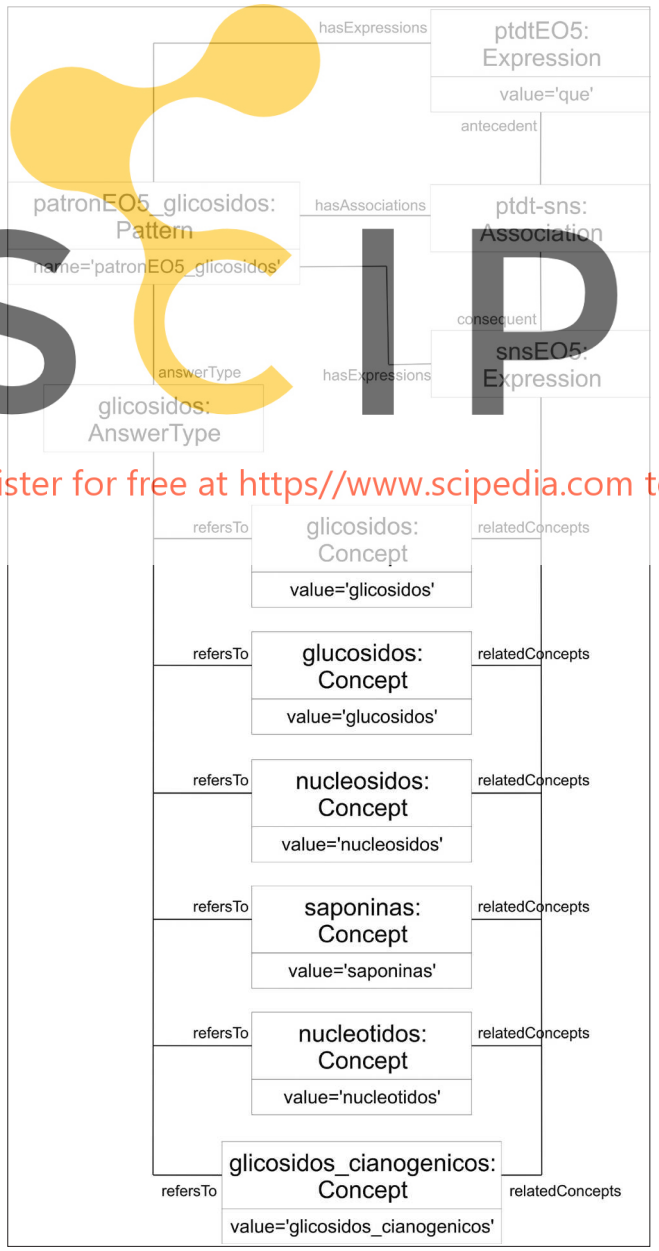

Figura 12. Ejemplo de modelo de patrón de respuesta adaptado al dominio por *Maraqa*

del sistema de BR de DA *AliQAn* sobre el mismo dominio de aplicación. Por ejemplo, para las primeras respuestas del sistema *baseline AliQAn* sobre el dominio agrícola *RCCA* fue de 28,8% y la del sistema de BR para el dominio *agrícola* fue de 58,3%. De esta manera nuestro sistema tuvo una mejora del 29,5% de precisión en las respuestas. Estos resultados demuestran la efectividad de nuestro método que supera la media del sistema *AliQAn* (43% de precisión) en todas sus actuaciones pasadas sobre dominios abiertos (**Roger** *et al.*, 2008).

Conclusiones

Se ha presentado el sistema BR *Maraqa*, realizado siguiendo el paradigma del desarrollo dirigido por modelos mediante la plataforma *Eclipse*. Permite adaptar fácilmente sistemas de BR en DA a dominios restringidos, integrando los recursos de conocimiento anteriormente utilizados con los nuevos recursos del DR, y generando el código fuente del sistema de BR adaptado a dicho dominio.

Para ello requiere como entrada la colección de documentos sobre la que se realizará la búsqueda de información, los recursos de conocimiento del DR independientemente del formato que tengan, y el código fuente de los patrones de preguntas y respuestas del sistema de BR en DA que se desea adaptar. A partir de estos recursos, genera la taxonomía de TRE refinada para el dominio en cuestión, y el código fuente de los patrones de preguntas y respuestas adaptados al dominio. Todo el proceso de funcionamiento de la herramienta ha sido ilustrado mediante un caso de estudio en el dominio agrícola.

Con *Maraqa* se supera el problema de la baja precisión alcanzada cuando un sistema de BR de DA trabaja en un dominio restringido (pasando de una precisión del 28,8% al 58,3%). Además se facilita la integración de los recursos de conocimiento de dicho dominio al sistema de BR, y todo ello a partir del corpus textual como punto de partida y fuente de información principal, en detrimento del uso de corpus de preguntas de otras propuestas, lo que consideramos poco realista en DR. La principal contribución es la adaptación automática de sistemas de BR existentes (específicamente de sus patrones de pregunta-respuesta y las taxonomías de TRE) a diferentes dominios restringidos, sin requerir ningún esfuerzo manual como el resto de propuestas previas existentes (**Sekine**; **Sudo**; **Nobata**, 2002; **Hovy**; **Hermjakob**; **Ravichandran**, 2002; **Metzler**; **Croft**, 2005; **Li**; **Roth**, 2006; **Ferrés**; **Rodríguez**, 2006; **Terol**; **Martínez-Barco**; **Palomar**, 2006; **Kosseim**; **Yousefi**, 2008; **Peñas** *et al.*, 2009).

Es posible la obtención de recursos de conocimiento (con información semántica) del dominio restringido, necesarios para aplicar *Maraqa*, a partir de la utilización de sistemas de alineamiento automático entre diferentes SOC's (**Soergel** *et al.*, 2004; **Van-Hage** *et al.*, 2010); lo que planteamos como un interesante trabajo futuro.

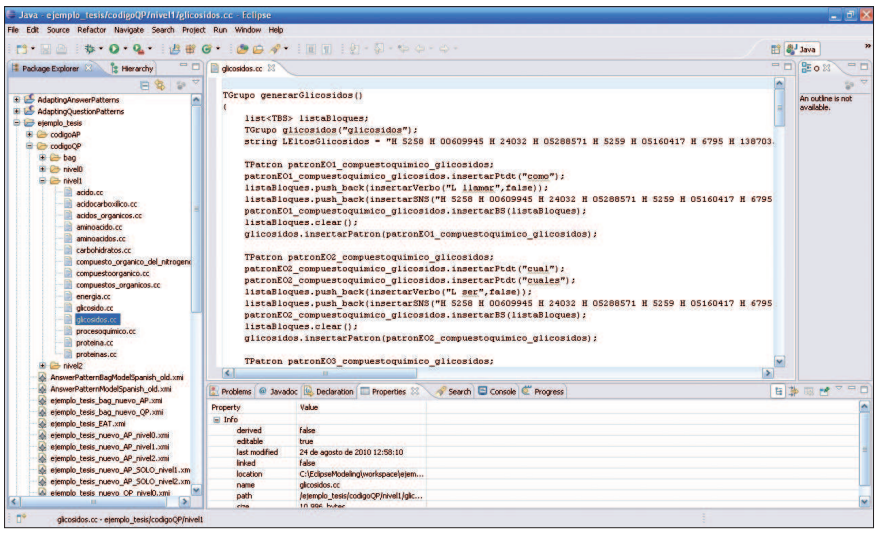

Figura 13. Ejemplo del código del patrón de pregunta adaptado al dominio por *Maraqa*

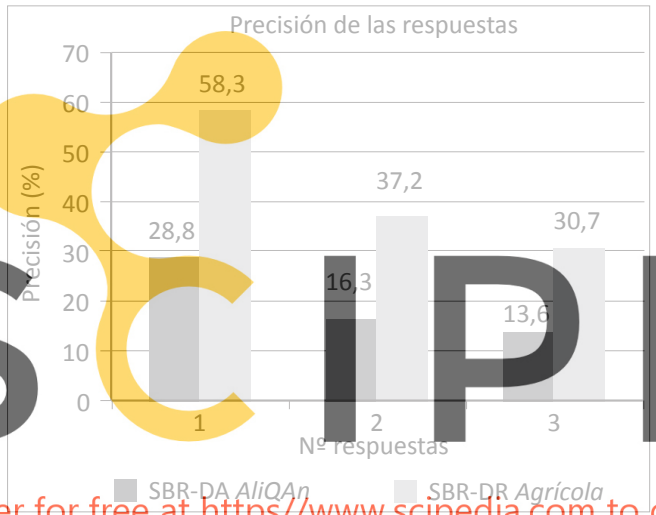

Figura 14. Precisión del sistema de BR para el dominio agrícola (desarrollado con *Maraqa*)

Agradecimientos

Esta investigación ha sido parcialmente financiada por el gobierno español bajo el proyecto *TIN2009-13391-C04-01*; y el gobierno valenciano bajo el proyecto *Prometeo/2009/119*. La investigación de **Katia Vila** ha sido financiada por una beca *MAEC-Aecid* del gobierno español.

Bibliografía

Acebo, S.; **Ageno, Alicia**; **Climent, Salvador**; **Farreres, Javier**; **Padró, Lluís**; **Placer, Roberto**; **Rodríguez, Horacio**; **Taulé, Mariona**; **Turmo, Jordi**. "*MACO*: *Morphological analyzer corpus-oriented*". *Esprit BRA-7315 Aquilex II. Working paper 31*, 1994.

Baeza-Yates, **Ricardo**; **Ribeiro-Neto**, **Berthier**. "Modern information retrieval". *ACM Press/Addison-Wesley*, 1999. ISBN: 978 0321416919

Bézivin, **Jean**. "On the unification power of models. Software and system modeling", 2005, v. 4, n. 2, pp. 171-188. *http://atlanmod.emn.fr/www/papers/OnTheUnificationPowerOfModels.pdf*

Ferrández, Antonio; **Palomar, Manuel**; **Moreno, Lidia**. "An empirical approach to Spanish anaphora resolution". *Machine translation*, 1999, v. 14, n. 3-4, pp. 191-216. *ftp://dlsi.ua.es/people/antonio/ART_MUL5.pdf*

Ferrés, Daniel; **Rodríguez, Horacio**. "Experiments adapting an open-domain question answering system to the geographical domain using scope-based resources". En: *MLQA'06: Procs of the Workshop on multilingual question answering. Association for Computational Linguistics*, 2006, pp 69-76. *http://dl.acm.org/citation.cfm?id=1708097.1708111*

Hodge, Gail. *Systems of knowledge organization for digital libraries: beyond traditional authority files*. The Digital Library Federation Council on Library and Information Resources, 2000. ISBN: 1887334769 *http://www.clir.org/pubs/reports/pub91/pub91.pdf*

Hovy, **Eduard**; **Hermjakob**, **Ulf**; **Ravichandran**, **Deepak**. "A question/answer typology with surface text patterns". En: *Procs of the 2nd intl conf on human language technology research*, Morgan Kaufmann Publishers Inc., 2002, pp. 247-251. *http://www.isi.edu/natural-language/projects/webclopedia/pubs/02hlt.pdf*

Katz, Boris. "From sentence processing to information access on the World wide web". *AAAI Spring symposium on natural language processing for the World wide web*, 1997, pp. 77-94. *http://aaaipress.org/Papers/Symposia/Spring/1997/SS-97-02/SS97-02-014.pdf*

Katz, Boris; **Borchardt, Gary**; **Felshin, Sue**. "Natural language annotations for question answering". En: *Procs of the 19th intl Flairs conf* (*Flairs* 2006), May 2006. *http://groups.csail.mit.edu/infolab/publications/FLAIRS0601KatzB.pdf*

Kleppe, Anneke; **Warmer, Jos**; **Bast, Wim**. *MDA explained. The practice and promise of the model driven architecture*. Addison Wesley, 2003. ISBN: 978 0321194428

Kosseim, Leila; **Yousefi, Jamileh**. "Improving the performance of question answering with semantically equivalent answer patterns". *Data and knowledge engineering*, 2008, v. 66, n. 1, pp. 53-67. *http://dx.doi.org/10.1016/j.datak.2007.07.010*

Li, Xin; **Roth, Dan**. "Learning question classifiers: the role of semantic information". *Natural language engineering*, 2006, v. 12, n. 3, pp. 229-249. *http://dx.doi.org/10.1017/S1351324905003955*

Mellor, Stephen J.; **Clark, Anthony N.**; **Futagami, Takao**. "Model-driven development - Guest editor's introduction". *IEEE software*, 2003, v. 20, n. 5, pp. 14-18. *http://dx.doi.org/10.1109/MS.2003.1231145*

Metzler, Donald; **Croft**, **W. Bruce**. "Analysis of statistical question classification for fact-based questions". *Informa-*

tion retrieval, 2005, v. 8, n. 3, pp. 481-504.
http://ciir.cs.umass.edu/pubfiles/ir-323.pdf
http://dx.doi.org/10.1007/s10791-005-6995-3

Mollá, Diego; **Vicedo, José-Luis**. "Question answering in restricted domains: an overview". *Computational linguistics*, 2007, v. 33, n. 1, pp. 41-61.
http://www.mitpressjournals.org/doi/pdf/10.1162/coli.2007.33.1.41
http://dx.doi.org/10.1162/coli.2007.33.1.41

Peñas, Anselmo; **Forner, Pamela**; **Sutcliffe, Richard**; **Rodrigo, Álvaro**; **Forascu, Corina**; **Alegria, Iñaki**; **Giampiccolo, Danilo**; **Moreau, Nicolas**; **Osenova, Petya**. "Overview of ResPubliQA 2009: question answering evaluation over European legislation". En: *Working notes of Cross language evaluation forum* (*CLEF*), 2009.
http://clef.isti.cnr.it/2009/working_notes/ResPubliQA-overview.pdf

Roger, Sandra; **Vila, Katia**; **Ferrández, Antonio**; **Pardiño, María**; **Gómez, José-Manuel**; **Puchol-Blasco, Marcel**; **Peral, Jesús**. "Using AliQAn in monolingual QA@Clef 2008". *9th Workshop of the Cross-language evaluation forum. Lecture notes in computer science*, 2008, v. 5706, pp. 333-336.
http://dx.doi.org/10.1007/978-3-642-04447-2_38

Russell, Stuart; **Norvig, Peter**. "Artificial intelligence: a modern approach". *Pearson education*, 2nd ed., 2003. ISBN: 978 0137903955

Sekine, Sstoshi; **Sudo, Kiyoshi**; **Nobata, Chikashi**. "Extended named entity hierarchy". En: *Procs of 3rd Intl conf on language resources and evaluation (LREC'02)*, 2002, pp. 1818-1824.
http://nlp.cs.nyu.edu/pubs/papers/sekine-lrec02.pdf

Soergel, Dagobert; **Lauser, Boris**; **Liang, Anita**; **Fisseha, Frehiwot**; **Keizer, Johannes**; **Katz, Stephen**. "Reengineering thesauri for new applications: the *Agrovoc* example". *Journal of digital information*, 2004, v. 4, n. 4.
http://journals.tdl.org/jodi/article/viewArticle/112/111

Terol, Rafael M.; **Martínez-Barco, Patricio**; **Palomar, Manuel**. "Aplicación de técnicas basadas en PLN al tratamiento de preguntas médicas en búsqueda de respuestas". *Procesamiento del lenguaje natural*, 2006, n. 36, pp. 17-24.
http://www.sepln.org/revistaSEPLN/revista/36/02.pdf

Van-Hage, Willem-Robert; **Sini, Margherita**; **Finch, Lori**; **Kolb, Hap**; **Schreiber, Guus**. "The OAEI food task: an analysis of a thesaurus alignment task". *Applied ontology*, 2010, v. 5, n. 1, pp. 1-28.
http://dx.doi.org/10.3233/AO-2010-0072

Vila, Katia. *Búsqueda de respuestas en dominios restringidos: aplicación sobre el dominio agrícola*. Tesis doctoral, Universidad de Alicante, 2010.
http://rua.ua.es/dspace/bitstream/10045/18329/1/tesis_vila.pdf

Vila, Katia; **Mazón, José-Norberto**; **Ferrández, Antonio**. "Model-driven development for adapting question answering systems to restricted domains". *Journal of research and practice in information technology*, May 2011, v. 43, n. 2, pp. 23-40.
http://dx.doi.org/10.1145/1966883.1966893